The Reproducible Research Movement: Crisis and Solutions

Victoria Stodden
Department of Statistics
Columbia University

Université d'Orléans
Orléans, France
November 6, 2012

Computation Emerging as Central to the Scientific Enterprise

- Enormous, and increasing, amounts of data collection:
  - CMS project at LHC: 300 "events" per second, 5.2M seconds of runtime per year, .5MB per event = 780TB/yr => several PB when data processed,
  - Sloan Digital Sky Survey: 8th data release (2010), 49.5TB,
  - quantitative revolution in social sciences social network data (Lazer et al., Science, 2009),
  - Science survey of peer reviewers: 340 regularly work with datasets >100GB; 119 regularly work with datasets >1TB (N=1700, Feb 11, 2011, p. 692)
- Massive simulations of the complete evolution of a physical system, systematically varying parameters,
- Deep intellectual contributions now encoded in software.

This research is supported by NSF award #1153384 and the Sloan Foundation.

Credibility Crisis

JASA June	Computational Articles	Code Publicly Available
1996	9 of 20	0%
2006	33 of 35	9%
2009	32 of 32	16%
2011	29 of 29	21%

Generally, data and code not made available at the time of publication, insufficient information captured in the publication for verification/replication of results.

➡ **A *Credibility Crisis***

Updating the Scientific Method

- Donoho and others argue that computation presents only a potential third branch of the scientific method:
- Branch 1 (deductive): mathematics, formal logic,
- Branch 2 (empirical): statistical analysis of controlled experiments,
- Branch 3? (computational): large scale simulations / data driven computational science.

The Ubiquity of Error

- The central motivation for the scientific method is to root out error:
  - Deductive branch: the well-defined concept of the proof,
  - Empirical branch: the machinery of hypothesis testing, structured communication of methods and protocols.
- Computational science as practiced today does not generate reliable knowledge. "breezy demos"
- See e.g. Ioannidis, "Why Most Published Research Findings are False," PLoS Med, 2005.

Emergent Efforts 1

- AMP 2011 "Reproducible Research: Tools and Strategies for Scientific Computing"
- AMP / ICIAM 2011 "Community Forum on Reproducible Research Policies"
- SIAM Geosciences 2011 "Reproducible and Open Source Software in the Geosciences"
- ENAR International Biometric Society 2011: Panel on Reproducible Research
- AAAS 2011: "The Digitization of Science: Reproducibility and Interdisciplinary Knowledge Transfer"
- SIAM CSE 2011: "Verifiable, Reproducible Computational Science"
- Yale 2009: Roundtable on Data and Code Sharing in the Computational Sciences
- ACM SIGMOD conferences
- NSF/OCI report on Grand Challenge Communities (Dec, 2010)
- IOM "Review of Omics-based Tests for Predicting Patient Outcomes in Clinical Trials"
- ...

Implementation Challenges

Interlocking set of incentives that influence scientific output:

- grant and funding agency requirements,
- journal and publication requirements,
- intellectual property constraints / patents and financial incentives,
- institutional expectations (hiring, promotion, awards),
- other legal policy: Congressional Acts, Whitehouse memoranda, collaboration requirements.

Funding Agency Policy

Funding Agency Policy

- NSF grant guidelines:

  "NSF ... expects investigators to share with other researchers, at no more than incremental cost and within a reasonable time, the data, samples, physical collections and other supporting materials created or gathered in the course of the work. It also encourages grantees to share software and inventions or otherwise act to make the innovations they embody widely useful and usable." (2005 and earlier)

- NSF peer-reviewed Data Management Plan (DMP), January 2011.

- NIH (2003): "The NIH endorses the sharing of final research data to serve these and other important scientific goals. The NIH expects and supports the timely release and sharing of final research data from NIH-supported studies for use by other researchers." (>$500,000, include data sharing plan)

NSF Data Management Plan

"Proposals submitted or due on or after January 18, 2011, must include a supplementary document of no more than two pages labeled 'Data Management Plan.' This supplementary document should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results." (http://www.nsf.gov/bfa/dias/policy/dmp.jsp)

NSF Data Management Plan

- No requirement or directives regarding data openness specifically.
- But, "Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing. Privileged or confidential information should be released only in a form that protects the privacy of individuals and subjects involved." (http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/aag_6.jsp#VID4)

Journal Policy

Experimental Setup

- Sample selection, computational research:
  - Select all journals from ISI classifications "Statistics & Probability," "Mathematical & Computational Biology," and "Multidisciplinary Sciences" (this includes Science and Nature).
  - Delete all journals that have ceased publication (5),
  - N = 170.
- Create dataset with ISI information (impact factor, citations, publisher) and supplement with publication policies as listed on journal websites, in June 2011 and June 2012.

Data Sharing Policy

	2011	2012	Change
Required as condition of publication, barring exceptions	18	19	1
Required but may not affect editorial decisions	3	10	7
Explicitly encouraged/addressed, may be reviewed and/or hosted	35	30	-5
Implied	0	5	5
No mention	114	106	-8

Code Sharing Policy

	2011	2012	Change
Required as condition of publication, barring exceptions	6	6	0
Required but may not affect editorial decisions	6	6	0
Explicitly encouraged/addressed, may be reviewed and/or hosted	17	21	4
Implied	0	3	3
No mention	141	134	-7

Supplemental Materials Policy

	2011	2012	Change
Required as condition of publication, barring exceptions	8	6	-2
Required but may not affect editorial decisions	7	10	3
Explicitly encouraged/addressed, may be reviewed and/or hosted	86	93	7
Implied	4	3	-1
No mention	64	58	-7

Findings

- Changemakers are journals with high impact factors.
- Progressive policies are not widespread, but being adopted rapidly.
- Close relationship between the existence of a supplemental materials policy and a data policy.
- Data and supplemental material policies appear to lead software policy.

Intellectual Property Constraints

Legal Barriers: Copyright

"To promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries." (U.S. Const. art. I, §8, cl. 8)

- Original expression of ideas falls under copyright *by default* (papers, code, figures, tables..)
- Copyright secures exclusive rights vested in the author to:
  - reproduce the work
  - prepare derivative works based upon the original

Exceptions and Limitations: Fair Use.

Responses Outside the Sciences 1: Open Source Software

- Software with licenses that communicate alternative terms of use to code developers, rather than the copyright default.

- Hundreds of open source software licenses:

  - GNU Public License (GPL)
  - (Modified) BSD License
  - MIT License
  - Apache 2.0 License
  - ... see http://www.opensource.org/licenses/alphabetical

Responses Outside the Sciences 2: Creative Commons

- Founded in 2001, by Stanford Law Professor Larry Lessig, MIT EECS Professor Hal Abelson, and advocate Eric Eldred.
- Adapts the Open Source Software approach to artistic and creative digital works.

Responses Outside the Sciences 2: Creative Commons

- Creative Commons provides a suite of licensing options for digital artistic works:
    - BY: if you use the work attribution must be provided,
    - NC: the work cannot be used for commercial purposes,
    - ND: no derivative works permitted,
    - SA: derivative works must carry the same license as the original

Response from Within the Sciences

The *Reproducible Research Standard* (*RRS*) (Stodden, 2009)

- A suite of license recommendations for computational science:
    - Release media components (text, figures) under CC BY,
    - Release code components under Modified BSD or similar,
    - Release data to public domain or attach attribution license.

➡ Remove copyright's barrier to reproducible research and,

➡ Realign the IP framework with longstanding scientific norms.

Winner of the Access to Knowledge Kaltura Award 2008

Copyright and Data

- Copyright adheres to raw facts in Europe.
- In the US raw facts are not copyrightable, but the original "selection and arrangement" of these facts is copyrightable. (Feist Publns Inc. v. Rural Tel. Serv. Co., 499 U.S. 340 (1991)).
  - ➡ Possibility of a residual copyright in data (attribution licensing or public domain certification).
  - ➡ Law doesn't match reality on the ground: What constitutes a "raw" fact anyway?

Legal Policy Barriers

Congress: America COMPETES

- America COMPETES Re-authorization (2011):
  - § 103: Interagency Public Access Committee:

    "coordinate Federal science agency research and policies related to the dissemination and long-term stewardship of the results of unclassified research, *including digital data* and peer-reviewed scholarly publications, supported wholly, or in part, by funding from the Federal science agencies." (emphasis added)

  - § 104: Federal Scientific Collections: OSTP "shall develop policies for the management and use of Federal scientific collections to improve the quality, organization, *access, including online access*, and long-term preservation of such collections for the benefit of the scientific enterprise." (emphasis added)

Whitehouse RFIs

- "Public Access to Peer-Reviewed Scholarly Publications Resulting From Federally Funded Research"
- "Public Access to Digital Data Resulting From Federally Funded Scientific Research"

  Comments were due January 12, 2012.

Bayh-Dole Act (1980)

- Promote the transfer of academic discoveries for commercial development, via licensing of patents (Tech Transfer Offices),
- Legislators blind to the coming digital revolution, and the impact on software patents and code release for reproducibility.
- Implications for science as a disruptor of openness norms:
    - patents => delay in revealing code, or closed code,
    - I assert Bilski => obfuscation of methods submitted for patents,
    - alters a scientist's incentives toward commercial ends, instead of the production of science as a public good.

Ideal Attributes of Tools

- Ability to verify computational results, with minimal burden to both the researcher and reviewer,
  - easy sharing of data and code (tracking of experiments, workflow, provenance),
  - easy re-use of data and code (download, licensing, executing).
- Incentives for code and data release through:
  - citation mechanisms,
  - supporting journal policies.

Emergent Efforts 2

Part of the Solution: Tools

- *Dissemination Platforms:*

RunMyCode.org	Madagascar	Open Science Framework
MLOSS.org	thedatahub.org	nanoHUB.org

- *Workflow Tracking and Research Environments:*

VisTrails	Kepler	CDE
Galaxy	GenePattern	Paper Mâché
Sumatra	Taverna	Pegasus

- *Embedded Publishing:*

Verifiable Computational Research	Sweave
Collage Authoring Environment	SHARE

Knowing what we know

- Tools vital in implementing verifiable science,
-

References

- Reproducible Research, Guest editor for Computing in Science and Engineering, July/August 2012.
- "The Scientific Method in Practice: Reproducibility in the Computational Sciences"
- "Enabling Reproducible Research: Open Licensing for Scientific Innovation"
- Reproducible Research: Tools and Strategies for Scientific Computing, July 2011
- Reproducible Research in Computational Science: What, Why and How, Community Forum, July 2011

available at http://www.stodden.net